AUTOMATIC DETECTION OF IRREGULAR ESSAYS

Yinghao Sun[1] and Vincent Kieftenbeld[2]
[1]Department of Psychology, College of Arts & Sciences, The Ohio State University [2]Pacific Metrics

Introduction

In educational testing, constructed-response questions have several advantages over traditional multiple choice questions. However, manually scoring such questions can be costly in terms of both time and expense. Fortunately, automated essay scoring systems provide an efficient, machine-based solution to scoring constructed-response questions. One open question in automated essay scoring is how to filter out irregular essays that do not meet basic requirements and should not be assigned a regular score, for example, essays that are off-topic. If irregular essays are not filtered out, validity of the scoring system can be impaired; for example, students may simply memorize a pre-written essay on an irrelevant topic yet submit it for a score.

Aims

The present project attempts to build a monitoring system, or a classifier, to detect irregular essays, through simple features, easily accessible machine learning algorithms, and ensemble learning methods.

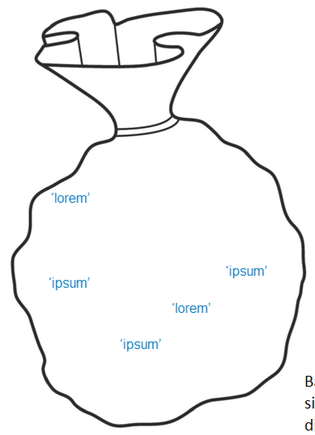
Methods

Dataset in this project came from a pilot test of a large scale assessment. There are 17 essay writing prompts and for each prompt a training set(approximately 1000 responses) and a test set (approximately 500 responses) are pre-defined. The averaged proportion of irregular essays for each prompt is around 20%.

In this project, detecting irregular essays is considered as a binary classification problem. Features were extracted with a bag-of-words approach, with term frequency, term frequency-inverse document frequency, and word/character n-grams. These high-dimensional features are selected using either truncated singular value decomposition or chi-squared statistics. Classification methods included k-Nearest Neighbors, naive Bayes, support vector machines, random forests, and ensemble learning such as stacked generalization.

Methods cont.

Ensemble learning is a way to combine and enhance the predictive performance of different types of classifiers. Ensemble learning methods in this project include majority voting, which assigns an essay a class label that agrees with the prediction of most classifiers. Winner-takes-all assigns a class label based on the prediction of the best classifier defined by cross-validation errors. Stacked generalization (or stacking) builds a meta-level classifier to combine predictions from different classifiers into a final prediction. Most classifiers were implemented in Python and with the scikit-learn package.



Bag-of-words is a simplified approach that discards grammar, order but multiplicity.

Results

Our results showed that ensembles decreased variance. Support vector machines outperformed other learning algorithms. Stacked generalization had marginal improvements in classification accuracy over other ensemble methods, achieving an average accuracy rate of 90%.

	Macroaverage (SD) on Test Samples			
	Accuracy	Recall	Precision	F1-score
Majority voting (binary)	.895 (.054)	.622 (.141)	.863 (.072)	.714 (.111)
Majority voting (probabilities)	.884 (.058)	.523 (.168)	.896 (.084)	.647 (.142)
Winner-takes-all	.897 (.054)	.666 (.138)	.834 (.084)	.732 (.104)
Stacking	.901 (.049)	.690 (.126)	.831 (.088)	.745 (.091)
NB	.862 (.057)	.479 (.162)	.815 (.113)	.585 (.132)
RC	.856 (.066)	.332(.241)	.934 (.082)	.446 (.250)
RF	.864 (.060)	.465 (.219)	.817 (.106)	.564 (.201)
SVM1	.888 (.053)	.601 (.131)	.837 (.088)	.693 (.109)
SVM2	.898 (.055)	.678 (.145)	.828 (.089)	.737 (.108)
SVM3	.898 (.053)	.667 (.136)	.834 (.080)	.733 (.103)
H1H2	.947 (.043)	.865 (.106)	.852 (.097)	.858 (.097)

Conclusion

Feature extraction and feature selection did play an important role in detection of irregular essays. Among learning algorithms, support vector machines outperformed other methods in our study. Among ensemble learning methods, stacked generalized outperformed winner-takes-all and majority voting, and showed marginal improvement over support vector machines. These results contribute to how automated scoring systems can selectively route responses for human review.

Bibliography

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. Natural Language Engineering, 12, 145-159.

Acknowledgements

I want to thank Vincent Kieftenbeld and Jie Li for their support and guidance, and CTB for the internship opportunity.

THE OHIO STATE UNIVERSITY